2007

The Use of R in Genetic Association Studies

MRC Epidemiology Unit, Cambridge, UK

jinghua.zhao@mrc-epid.cam.ac.uk

10 August 2007

MRC | Epidemiology Unit

Outline

- Some background
- Genetic Association
  - Principles and Problems
  - Examples
- R implementation
  - What are available in R
  - Features and Examples
- General issues

Genetic Epidemiology

- *The study of the joint action of genes and environmental factors in causing disease in human population and their patterns of inheritance in families*
  - Thomas DC. *Statistical Methods in Genetic Epidemiology.* Oxford, Oxford University Press 2004
- *A science that deals with the aetiology, distribution and control of disease in groups of relatives and with inherited causes of disease in population… genetic epidemiology should be broadly defined to embrace all aspects of population genetics except evolution, including phenotypes, gene-environmental interactions, and modes of transmission related to health or to location of disease genes, or requiring methods of analysis developed for genetic determinants of disease… other disciplines are subsumed by this definition, including ecogenetics, behaviour genetics, and demographic genetics … molecular epidemiology as it relates to inherited risk factors is contained within genetic epidemiology*
  - Morton NE. *Annu Rev Genet* **1993;** 27: 523-38

Some Facts about Human Genome

- 23 (22 autosomal+1 sex) pairs of chromosomes
- 3x10^9 DNA base pairs (bp), or 50,000~100,000 genes if 30,000bp per gene
- Human Genome Projects and HAPMAP projects (Couzin. *Science* 296:1391-3, 2002; 310:601, 2005; Couzin & Kaiser. *Science* 316:822,2007)
  - http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
  - http://www.hapmap.org

A Paradigm of Gene-Disease Association Study

Lander & Schork. *Science* 265:2037-2048,1994

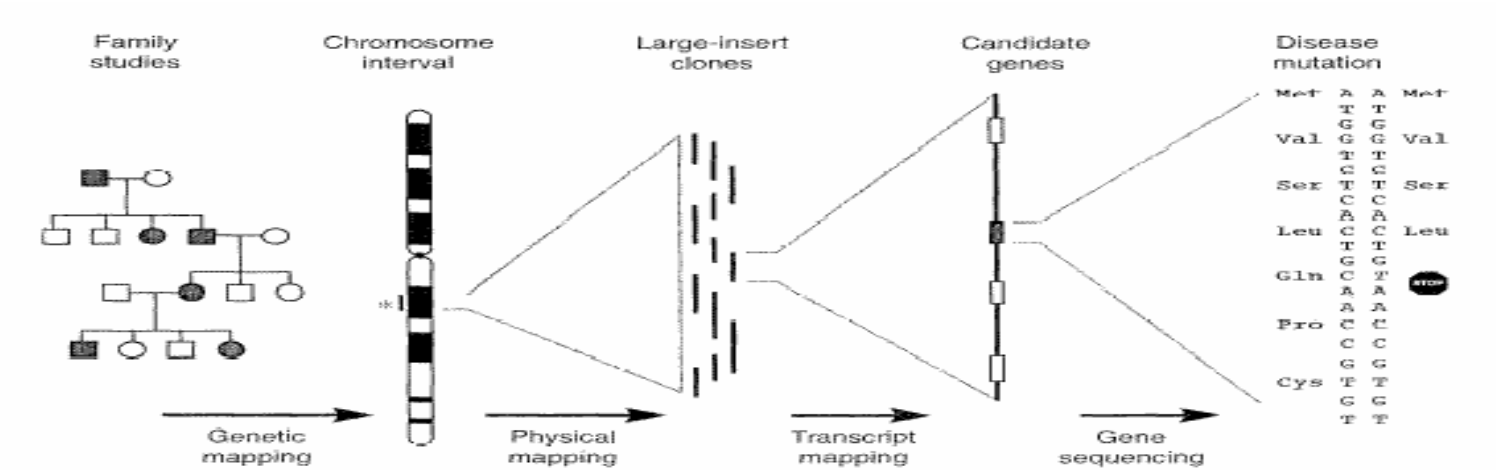


Fig. 1. Steps in positional cloning. Positioning of disease loci to chromosomal regions with genetic markers has become increasingly straightforward, particularly given the recent release of the Généthon genetic map containing 5264 markers (*17*). However, identification and evaluation of the genes within the implicated region remains a major stumbling block.

Schuler GD, et al.(1996) Science 274:540-546, 1996

Genetic Association

- Traditional (e.g. segregation and linkage) methods in human gene-trait study gradually move towards association study.
- This could either be population-based or family-based sample
- Large-scale studies are required to maintain statistical power, and it is also necessary to maintain a good coverage of the genome
- Both disease and trait are subject to scrutiny
  - WTCCC
  - The anthropometric consortium (height, body-mass index (BMI, kg/m^2)
- It is hugely important but difficult and costly
- A better integration of statisticians, computer professionals and other researchers is required

Three Initiatives

- The hapmap project, a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals.
  - http://www.hapmap.org
- *The Wellcome Trust Case Control Consortium (WTCCC) was formed with a view to exploring the utility, design and analyses of GWA studies. It brought together over 50 research groups from the UK that are active in researching the genetics of common human diseases, with expertise ranging from clinical, through genotyping, to informatics and statistical analysis.*
  - http://www.wtccc.org
- The genetic association information network (GAIN, a public-private partnership of the Foundation for the National Institutes of Health, Inc., which include corporations, private foundations, advocacy groups, concerned individuals, and the National Institutes of Health. This initiative will take the next step in the search to understand the genetic factors influencing risk for complex human diseases.
  - http://www.fnih.org/GAIN2/home_new.shtml,

Genome-wide Association Study (GWAS)

- Advance in genotyping and sequence technology make it possible to have large number of DNA variants called single-nucleotide polymorphisms (SNPs)
- Example include Perlegen 250k, Affymetrics 500k, Illumina 317k, GeneChips
- Population-based or family-based samples
- 10,000 individuals each with 1 million SNPs has a total of genotypes for 1GB storage if single byte variables are used. However, the storage could be reduced if data are organised by chromosomes. It is possible to pack SNP genotypes into 2 bits

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

A collaborative study of Bipolar disorder, Coronary artery disease, Hypertension, Type-1 Diabetes, Type-2 Diabetes, Crohn's disease, Rheumatic Arthritis

Statistical/Computational Perspectives

- Study designs
- Hardy-Weinberg equilibrium tests, quality control, statistical modelling
- Meta-analysis
- Features/structure via data mining or machine learning
- Evolutionary theory, in which current data are seen as snapshots of human evolutionary history
- Bayesian methods

Examples

- EPIC-Norfolk study of obesity (Affymetrix 500k and Illumina 317 GeneChips), 25~30 considered as overweight and >=30 as obesity, which involves ~7,000 individuals in a two-stage design; EPIC study of gene-environmental interaction of type-2 diabetes (10,000 incident cases plus controls)
  - http://www.srl.cam.ac.uk/epic/
- Collaborative studies with BC1958 cohort, GlaxoSmithKline and other institutions

A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity

Timothy M. Frayling,[1,2]* Nicholas J. Timpson,[3,4]* Michael N. Weedon,[1,2]* Eleftheria Zeggini,[3,5]* Rachel M. Freathy,[1,2] Cecilia M. Lindgren,[3,5] John R. B. Perry,[1,2] Katherine S. Elliott,[3] Hana Lango,[1,2] Nigel W. Rayner,[3,5] Beverley Shields,[2] Lorna W. Harries,[2] Jeffrey C. Barrett,[3] Sian Ellard,[2,6] Christopher J. Groves,[5] Bridget Knight,[2] Ann-Marie Patch,[2,6] Andrew R. Ness,[7] Shah Ebrahim,[8] Debbie A. Lawlor,[9] Susan M. Ring,[9] Yoav Ben-Shlomo,[9] Marjo-Riitta Jarvelin,[10,11] Ulla Sovio,[10,11] Amanda J. Bennett,[5] David Melzer,[1,12] Luigi Ferrucci,[13] Ruth J. F. Loos,[14] Inês Barroso,[15] Nicholas J. Wareham,[14] Fredrik Karpe,[5] Katharine R. Owen,[5] Lon R. Cardon,[3] Mark Walker,[16] Graham A. Hitman,[17] Colin N. A. Palmer,[18] Alex S. F. Doney,[19] Andrew D. Morris,[19] George Davey Smith,[4] The Wellcome Trust Case Control Consortium,† Andrew T. Hattersley,[1,2]‡§ Mark I. McCarthy[3,5]‡

Analytical Issues

- General issues
  - Balding. *Nat. Rev. Genet*, **7**, 781-791, 2006
  - Elston & Anne Spence. *Stat Med.*, **25**, 3049-3080, 2006
- Specific issues
  - Haplotype-tagging and imputation
  - Multiple-testing, but unlike gene-expression data to use q-value/pFDR
  - Population stratification and outlier detection
  - Genotyping error, calling algorithms

Computer Implementations

- Algorithms evolution span over ~50 years
- Standalone software programs (Pascal, Fortran, C/C++,…) (http://inkage.rockefeller.edu)
- Increasingly limited and requires marriage with established systems (commercial and free, e.g. SAS/Stata/S-PLUS and R).
- In particular
  - The call for integration is reminiscent of the development of Linux system
  - R is a general computing environment for both practitioners and programmers. Standard statistical and genetic analyses can be conducted in a reliable and unified fashion. It makes efficient statistical modelling possible.
  - It is accessible and free
  - It is easy to extend and a large number of packages maintained by a large and dedicated team
  - Zhao & Tan. *Hum. Genomics*, **2**, 258-265, 2006

SAS

- An established systems (http://www.sas.com, , http://en.wikipedia.org/wiki/SAS) for many fields of research
- Database support such as Oracle, MySQL and facility such as ODBC, data visualisation, fantastic data subsetting facility which offers better solution than INNER JOIN, i.e.

```
PROC SQL;
   CREATE TABLE b AS SELECT * FROM  d1 WHERE rsn NOT IN
    (SELECT rsn FROM d2);
QUIT;
```

- Modelling facilities, statistics, operations research
- Not without problems (e.g. sorting and long-format data) and better used in conjunction with other systems such as Stata (http://ww.stata.com) and R (http://www.r-project.org)

Stata

The most popular software for epidemiologists, and researchers in other fields such as econometrics. Its facilities range from basic data management, computer graphics, programming, to methods for complex survey, longitudinal data analysis and multilevel models. Beside its simple but flexible syntax and comprehensive on-line documentation, it has many unique features such as frequency, probability and analytic weights.

- library(foreign) functions read.dta/write.dta
- Zhao & Tan. *Curr Bioinformatics, 2006*

R

- Generic packages as given in ctv for genetics
  - genetics (Warnes. *R News*, **3**, 9-13, 2003)
  - haplo.stats (Lake et al. *Hum. Hered.*, **55**, 56-65, 2003)
  - gap (Zhao J. Stat Soft)

- Packages for genomewide association studies (GWAS)
  - SNPassoc (Gonzalez et al. *Bioinformatics*, **23**, 644-645, 2007)
  - GenABEL (Alchenko et al. *Bioinformatics*, **23**, 1294-1296, 2007)
  - SNPmatrix (Clayton and Leung, *Hum. Hered.*, **64**, 45-51, 2007)

Features

- Generic -- genetics
- Problem specific – haplo.stats
- An incremental collection of available codes -- gap
  - Design: Risch & Merikangas *Science*, 1996; Cai & Zeng. Biometrics. 2004; Skol et al. *Nat Genet.* 2006
  - Analysis: Zhao et al. *Hum Hered*. 2000; Zapata et al. *Ann Hum Genet. 2001;* Schaid et al. *Am J Hum Genet* 2002; Zaykin et al. Hum Hered 2002; Wacholder et al. *J Natl Cancer Inst.* 2004; Wakefield. *Am J Hum Genet* 2007

GWASware

- Hybrid approaches
- Analytic aspects ranging from QCs to summary statistics, multi-locus model and gene-environmental interactions
- Unique features from each package
- Largely focused on SNPs

SNPassoc

```
library(SNPassoc)
snps <- setupSNP(g4, colSNPs=5:3961, sort = TRUE, info=g4pos)
snps1 <- snps[stage1,]
ans1 <- WGassociation(cc~b58cregion, data=snps1, model="log")
```

GenABEL

```
data(ge)
a <- ibs(data=ge,ids=c(1:10),snps=c(1:1000))
a
# compute IBS based on a random sample of 1000 autosomal marker
a <-
    ibs(ge,snps=sample(ge@gtdata@snpnames[gec@gtdata@chromosome!="
    X"],1000,replace=FALSE))
mds <- cmdscale(as.dist(1-a))
plot(mds)
# identify smaller cluster of outliers
km <- kmeans(mds,centers=2,nstart=1000)
cl1 <- names(which(km$cluster==1))
cl2 <- names(which(km$cluster==2))
if (length(cl1) > length(cl2)) cl1 <- cl2;
cl1
# PAINT THE OUTLIERS IN RED
points(mds[cl1,],pch=19,col="red")
```

HAPMAP Data

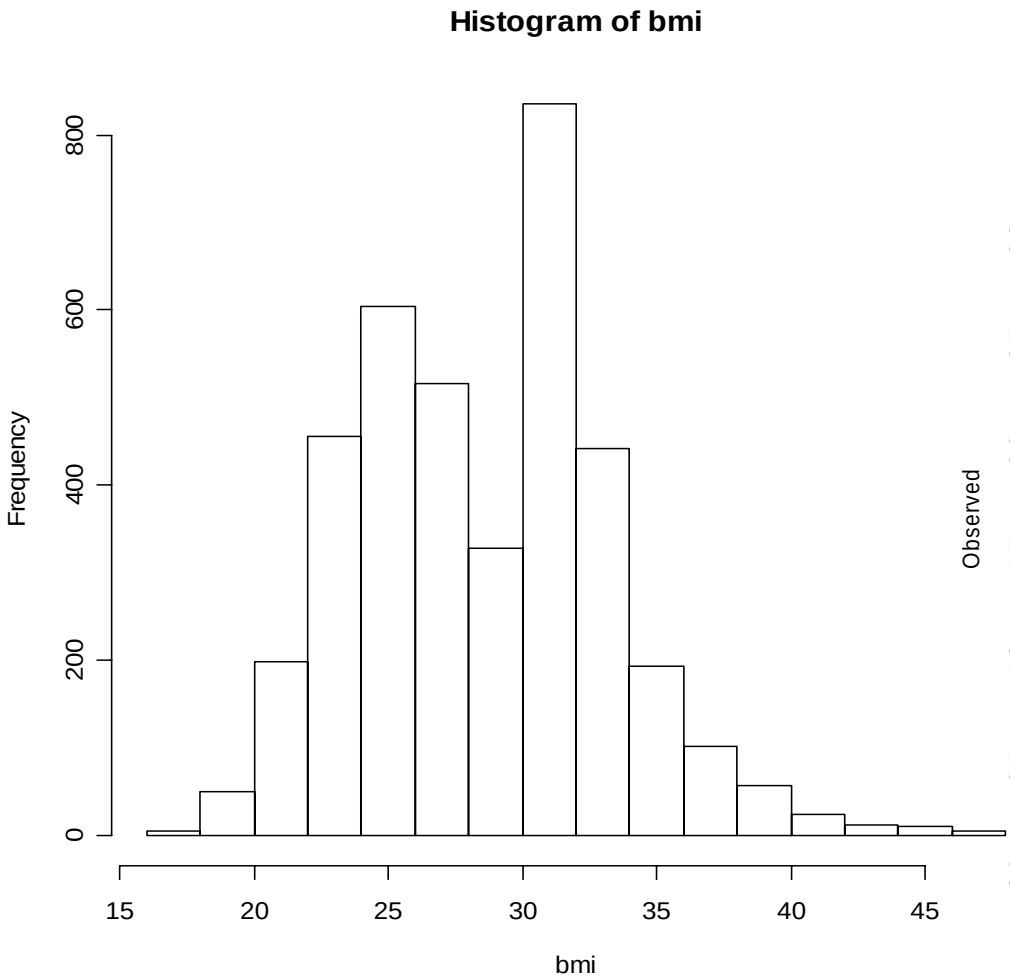
```
library(snpMatrix)
baseurl <- "http://www.hapmap.org/genotypes/latest/fwd_strand/non-redundant/";
CEU <- scan("CEU.lst",what="")
hapmap <- function(i=1)
{
  hapmap <- paste(baseurl,CEU[i],sep="")
  result <- read.HapMap.data(hapmap)
  sum <- summary(result$snp.data)
  invisible(list(snp.support=result$snp.support,snp.sum=sum[is.finite(sum$z.HWE),]))
}
for (i in 1:24)
{
  if (i==1)
  {  hapmap1 <- hapmap(1)
     snp.support <- hapmap1$snp.support
     snp.sum <- hapmap1$snp.sum
  } else {
     hapmapi <- hapmap(i)
     snp.support <- rbind(snp.support,hapmapi$snp.support)
     snp.sum <- rbind(snp.sum,hapmapi$snp.sum)
  }
}
```

Meta-analysis

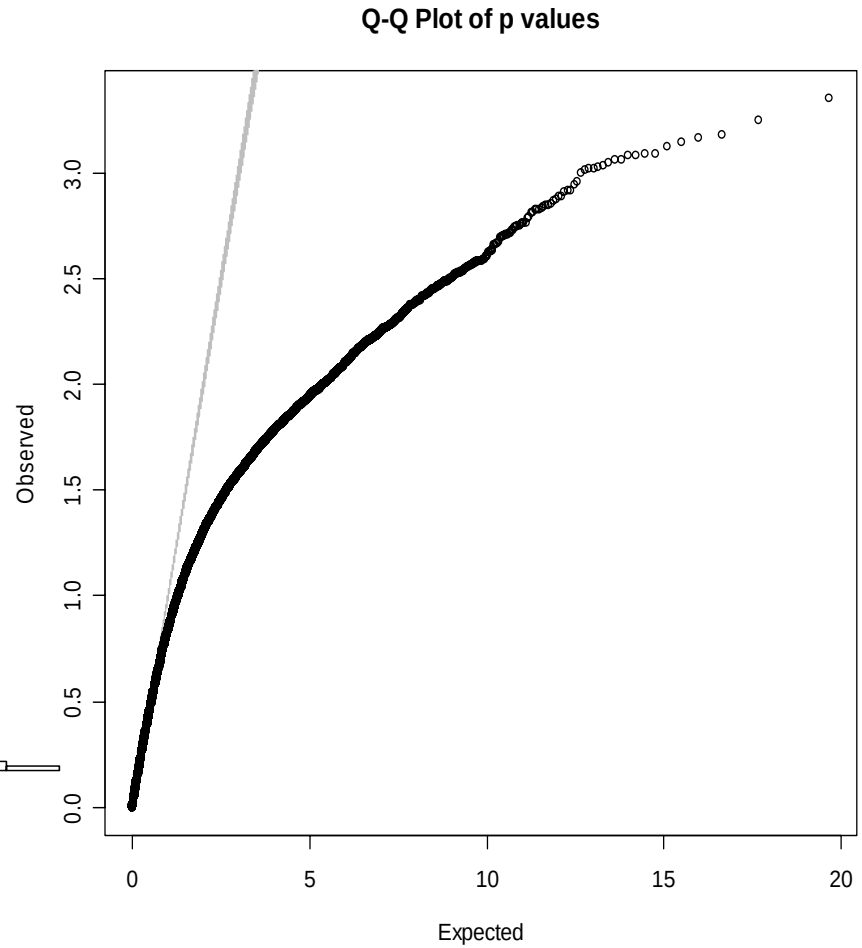
```
library(foreign)
meta5 <- read.dta("meta5.dta")
attach(meta5)
library(meta)
by(meta5,locus,function(x)
   metagen(b,se,data=x))
```

Results/Problems

- It has been suggested %10 SNPs could be carried over to stage 2
- But 149 SNPs were selected for fast-tracking
- Still difficult with continuous traits such as BMI, meta-analysed and currently in joint efforts with Oxford, Harvard-MIT Broad Inst, FUSION, …



The BMI distribution in the EPIC-Norfolk study of obesity. The case-cohort sample is a combination of the sub-cohort sample and case sample which is truncated from the whole EPIC cohort at BMI=30.



Q-Q plot of p values from PROC QLIM

Features not Well-appreciated by Most Analyses

- Probability weighting, similar to SAS but not Stata
- Bayesian methods as implemented in R2WinBUGS, MCMCpack, mcmc, or BRugs, etc

A Comparison with Standalone Programs

- Most implementations (e.g. PLINK: Purcell et al. *Am J Hum Genet,* 2007; EIGENSTRAT. Price et al. *Nat Genet.* 2006; SNPTEST: Marchini et al. *Nat Genet* 2007) have consolidated forms in R
  - Goodness of fit test (HWE) and association testing (Hotelling's T^2, etc)
  - MVA (regression, cluster, correspondence, principal component, Latent class, multidimensional scaling) and confirmatory analysis
  - Graphics
- It is preferable to have independent compression/decompression and data-handling algorithms implemented in R
- The primary driving force is that genetic data increasingly integrated, so that established systems are required and implementation involves a large number of collaborators

Other Aspects and Prospects

- Family-based studies are not covered here
- Given the advance in technologies the scale of data will be even larger, and more data and results are publicly available together with other types of data
- Routine but comprehensive analysis can be done with established packages
- For most, it is largely untenable to see R programming as purely implementation but to facilitate development of statistical models and computational algorithms and tools towards better exposition of data and understanding of the underlying socio-biological processes; R will play a irreplaceable and increasingly bigger role

Acknowledgements

- Dianne Cook and useR 2007 program committee
- Special thanks to colleagues who contribute their codes and provide support
- Institutions
  - MRC Epidemiology Unit
  - The Wellcome Trust Sanger Institute
  - The University of Cambridge

NHS

MRC | Epidemiology Unit

UNIVERSITY OF CAMBRIDGE

Institute of Metabolic Science (http://www.ims.cam.ac.uk)